

PROCEEDINGS OF SPIE

Assurance and Security for AI-enabled Systems 2026

**Joshua D. Harguess
Nathaniel D. Bastian
Chris M. Ward**
Editors

**27–29 April 2026
National Harbor, Maryland, United States**

Sponsored and Published by
SPIE

Volume 14046

Proceedings of SPIE 0277-786X, V. 14046

SPIE is an international society advancing an interdisciplinary approach to the science and application of light.

The papers in this volume were part of the technical conference cited on the cover and title page. Papers were selected and subject to review by the editors and conference program committee. Some conference presentations may not be available for publication. Additional papers and presentation recordings may be available online in the SPIE Digital Library at SPIDigitalLibrary.org.

The papers reflect the work and thoughts of the authors and are published herein as submitted. The publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Please use the following format to cite material from these proceedings:

Author(s), "Title of Paper," in *Assurance and Security for AI-enabled Systems 2026*, edited by Joshua D. Harguess, Nathaniel D. Bastian, Chris M. Ward, Proc. of SPIE 14046, Seven-digit Article CID Number (DD/MM/YYYY); (DOI URL).

ISSN: 0277-786X

ISSN: 1996-756X (electronic)

ISBN: 9798902320487

ISBN: 9798902320494 (electronic)

Published by

SPIE

P.O. Box 10, Bellingham, Washington 98227-0010 USA

Telephone +1 360 676 3290 (Pacific Time)

SPIE.org

Copyright © 2026 Society of Photo-Optical Instrumentation Engineers (SPIE).

Copying of material in this book for internal or personal use, or for the internal or personal use of specific clients, beyond the fair use provisions granted by the U.S. Copyright Law is authorized by SPIE subject to payment of fees. To obtain permission to use and share articles in this volume, visit Copyright Clearance Center at copyright.com. Other copying for republication, resale, advertising or promotion, or any form of systematic or multiple reproduction of any material in this book is prohibited except with permission in writing from the publisher.

Printed in the United States of America by Curran Associates, Inc., under license from SPIE.

Publication of record for individual papers is online in the SPIE Digital Library.

**SPIE. DIGITAL
LIBRARY**

SPIDigitalLibrary.org

Paper Numbering: A unique citation identifier (CID) number is assigned to each article in the Proceedings of SPIE at the time of publication. Utilization of CIDs allows articles to be fully citable as soon as they are published online, and connects the same identifier to all online and print versions of the publication. SPIE uses a seven-digit CID article numbering system structured as follows:

- The first five digits correspond to the SPIE volume number.
- The last two digits indicate publication order within the volume using a Base 36 numbering system employing both numerals and letters. These two-number sets start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B ... 0Z, followed by 10-1Z, 20-2Z, etc. The CID Number appears on each page of the manuscript.

Contents

vii *Conference Committee*

ADVERSARIAL AI I

- 14046 03 **Optical adversaries: assessing robustness of test time adaptation against light-based adversarial perturbation** [14046-1]
- 14046 04 **3D reconstruction and neural rendering for adversarial machine learning** [14046-2]
- 14046 05 **Evaluating natural and generative trojans in visual object detection** [14046-41]
- 14046 06 **A comparative analysis of UAS computer vision systems in physically realizable adversarial AI scenarios** [14046-4]

ADVERSARIAL AI II

- 14046 07 **Real-world adversarial camouflages using programmable E-ink panels for robot detection evasion** [14046-5]
- 14046 08 **Bypassing fabrication: physical adversarial attacks using everyday items** [14046-6]
- 14046 09 **Giving AI a headache: acoustic adversarial attacks to computer vision applications** [14046-7]

AI AGENT AND LLM SECURITY

- 14046 0A **MCP safety audit: LLMs with the model context protocol allow major security exploits** [14046-8]

AI MISSION AUTONOMY

- 14046 0C **A causal explainability framework for trustworthy mission autonomy** [14046-12]
- 14046 0D **Measurable mission impact of adversarial AI** [14046-13]
- 14046 0E **DP-TwoLevel: two-stage gradient subspace learning for differentially private federated learning** [14046-14]

AI GOVERNANCE I

- 14046 OF **SPARROW: a simulation platform for evaluating adversarial resilience in autonomous aerial systems** [14046-11]
- 14046 OG **A hybrid threat modeling and risk validation framework for mission-critical AI systems** [14046-16]
- 14046 OH **Operationalizing enterprise AI security: an open-source framework for AI cyber risk management** [14046-17]
- 14046 OI **Quantifying the unseen: a framework for separability assessment through reconstruction and retraining** [14046-22]

TEVV FOR AI SYSTEMS

- 14046 OJ **AI-CORR: artificial intelligence-cyber operations research range** [14046-19]
- 14046 OK **Behavioral fingerprinting of neural networks via topology-guided fuzzing** [14046-20]
- 14046 OL **Verification of deep neural networks using latent variable generative models** [14046-21]
- 14046 OM **From detection to correction: uncertainty-based drift detection for ML models for motion imagery** [14046-23]

CYBER DEFENSE AND RED TEAMING

- 14046 ON **DC-IDS: a dynamic continual learning intrusion detection system with uncertainty-aware adaptation** [14046-24]
- 14046 OO **VulnSieve: a structured knowledge preprocessing engine for agentic AI penetration testing** [14046-25]
- 14046 OR **A red teaming framework for evaluating robustness of AI-enabled security orchestration, automation, and response systems** [14046-28]

AI MODEL HARDENING

- 14046 OT **Fast and generalized robustness via distillation of energy-based models trained with equilibrium propagation** [14046-29]
- 14046 OU **Assessing the potential for catastrophic failure in dynamic post-training quantization** [14046-30]
- 14046 OV **Corrective unlearning through targeted node ablation** [14046-31]

AI GOVERNANCE II

- 14046 0W **AI toolkit for federal agencies** [14046-18]
- 14046 0Y **GaleStorm: deterministic, privacy-preserving AI oversight for digital assets and healthcare compliance** [14046-33]

EMERGING TOPICS IN AI SECURITY

- 14046 10 **An exploration of ethical AI design issues in AI-enabled command and control systems** [14046-36]
- 14046 11 **Model fingerprinting for secured AI supply chain** [14046-38]
- 14046 12 **Improving semantic coverage in AI test and evaluation through generative augmentation** [14046-39]
- 14046 13 **Metacognition AI framework for effective fusion and reasoning over media forensics analytics** [14046-40]

POSTER SESSION

- 14046 14 **The challenge of autonomous cyber defense** [14046-42]
- 14046 15 **Semantic similarity deficits in neural network critical learning periods** [14046-43]
- 14046 16 **Many metrics, one truth? Combining heuristics to detect hallucinations in LLMs** [14046-44]
- 14046 17 **Query-efficient finetuning to improve adversarial attack transferability** [14046-45]
- 14046 18 **Responsible design of research focused artificial-intelligence-enabled systems** [14046-46]
- 14046 19 **Designing image augmentations to bridge the synthetic-to-real data gap** [14046-47]
- 14046 1A **Entropy-weighted multilayer attention for token-level attribution in autoregressive language models** [14046-48]
- 14046 1B **Operationalizing AI assurance: translating maturity and risk frameworks into measurable practice** [14046-49]
- 14046 1C **Synthetic battlefields: benchmarking scene-level AI deception and developing resilient detectors for defense intelligence** [14046-50]
- 14046 1D **AITE: a modular and extensible framework for robust AI testing and evaluation** [14046-52]